geo✗walk

Title: Full Report for geoXwalk Phase II project.

Synopsis: Full report for the Phase II goeXwalk project covering methodology, activities and future priorities.

Author: EDINA, GeoData Services

Date: 03 September 2003

Version: 1.2

Status: Final

Authorised:

EDiNA

H|D|S
History Data Service

Background to project

The geoXwalk Phase II project was a JISC funded one year development project aimed at developing a demonstrator gazetteer server and service for the purposes of enhancing geographic searching within the JISC Information Environment. It followed on from a successful Phase I scoping study which outlined the requirements and stakeholders for such a service. The project was a joint one between EDINA, University of Edinburgh and the History Data Service, University of Essex.

The principal purpose of geoXwalk is to provide a shared service within the JISC Information Environment (IE) that can underpin geographic searching. The rationale behind the project is that there is currently no unified entry point to assist in geographic searching within the existing academic network as each information provider/service adopts different geographic coding conventions (some use postcodes, others placenames, some grid references etc.). geoXwalk is designed to make geographic searching transparent by 'crosswalking' these different geographies and is analagous to a shared terminology service.

geoXwalk provides more than just a simple lookup facility however as every geographic feature stored in the gazetteer has its detailed geometry stored with it (i.e. a city would be stored as a polygonal footprint (co-ordinate list), a river as a linear footprint etc.). Holding the geometry as an integral attribute of the feature enables complex spatial searching based on relationships between features e.g. is feature A within a distance of feature B?; what features are contained within feature C?; what features does feature D intersect? and so on.

The ability to derive the relationships between features implicitly by geometric computation is significant and provides more flexibility in the results than can be ascertained by simple lookups based on hierarchical thesauri methods, as is traditional in gazetteers. Furthermore, geography in the UK is very complex and geographic boundaries in particular do not always nest, for example, postcode geography does not nest with census geography.

geoXwalk obviates the problem of variable geographic naming by coding geographic features based on a persistent and consistent coding convention - national grid references. Again, by use of the implicit relationships that can be inferred from their geometries, it is possible to 'crosswalk' place names to: postcodes …or electoral wards… or health authorities etc…

As part of the project, the need for a geoparser was identified. That is, software that can take a document/resource that contains placenames and automatically identify their occurrence. Having identified a placename as such, the next logical step is to compare that against entries in the gazetteer which provides a means to access its 'alternate' geographies. For example, the placename 'Knowsley' could be resolved as parish code 'BX003' or grid reference 340900, 392300 - 347217, 397660. This methodology provides a means to explicitly georeference (i.e. attach a grid reference) implicitly georeferenced material (such as 'Knowsley'). The result is that more powerful geographical based search strategies can be applied e.g. find me all documents about Gaelic songs that do not reference the Western Isles or, find me images of towns along the river Tweed etc.

Methodology

The primary objective of the project was to develop a working gazetteer server and service that could demonstrate geographical querying within a machine2machine context. As a consequence a range of key technical issues had to be addressed.

Technical

The project involved a range of technical aspects involving which are reported more fully on in the supporting separate reports. The key technical aspects include:

- Assessment of the Alexandria Digital Library Content Standard, Feature Thesaurus and Query Protocol for supporting the geoXwalk service;

- Data enhancements and methodologies for improving the spatial accuracy of extant gazetteer sources that have seeded geoXwalk;

- The utility and practicability of supporting other query protocols, specifically Z39.50.

User engagement/Stakeholder involvement

It was decided to seek the opinions of the stakeholder community on the functionality of and the strategy for developing geoXwalk via a variety of methods. An analysis of approaches employed within usability and user testing was undertaken early on. It was decided that the following methods would be used to engage with users:

- direct communication via a JISCmail list;
- demonstrations;
- survey-based evaluation via paper questionnaires and focus group discussion sessions;
- communication via journal articles and newsletters;
- conference exhibitions;
- meetings with stakeholders, including a stakeholder workshop

A separate report, 'geoXwalk Evaluation Report' details stakeholder feedback.

Activities (including evaluation approaches)

Project management

The project has been managed at EDINA and has entailed regular meetings, teleconferences and email correspondence with HDS at Essex. Requisite project plans and bi-annual reports have been submitted to the JISC office as per the project timetable.

User engagement/Stakeholder involvement

At the very start of phase II a list of stakeholders was identified and a JISCmail list established for easy communication. Following this, the six methods taken to involve users and stakeholders in the development of geoXwalk were put into practice through the following activities:

- new developments were announced and opinions sought via the JISCmail list;
- the service was demonstrated at a workshop held at the University of Essex on 30 January 2003;
- the January workshop also asked relevant stakeholders to comment on the strategies to be taken to develop further geoXwalk and to integrate it into other services;
- the January workshop included survey-based evaluation of the design and potential functionality of geoXwalk via paper-based questionnaires;
- journal articles in Burisa, EDINA Newsline, WYVERN and UKDAtabytes were published in order to raise awareness of geoXwalk more widely;
- geoXwalk was represented at the GISRUK 2003 conference, via the EDINA stand; representatives from the service also attended the AGI conference in 2002 in order to raise awareness of its functionality and engage with some of the potential users.

Gazetteer population

The gazetteer database has been successfully seeded with near contemporary data sources. Content includes both extensive geographical coverage (GB) and thematic content. A complete listing is given in Appendix A of the separate report entitled: 'Alexandria Digital Library Content Standard and Data Loading & Processing Issues Report.' Over 330,000 individual objects many with complex spatial footprints have been added to the gazetteer.

Discussions on populating the gazetteer took place at the stakeholder workshop on 30 January 2003. Stakeholders discussed adding different data to the gazetteer, such as canals, roads and historical data. The latter will form a key task for Phase III of the project.

Geoparser

Software for the semi-automatic geographical indexing of existing implicitly geographically referenced resources has been developed (referred to as a 'geopaser'). Work on the geoparser originally formed only a small part of the overall geoXwalk project but due to stakeholder interest and the generic utility of a geoparser, subsequently blossomed into something more important. A reference web based front end to the geoparsing software has also been developed to assist in user review of the parsing results.

The geoparser was introduced to stakeholders at a workshop on 30 January 2003. The stakeholders were invited to complete a survey-based questionnaire and to participate in focus group discussion sessions, which resulted in:

- the evaluation of the geoparser via a survey-based questionnaire focusing on the methods used, design, parser output and confidence in results;
- focus group discussion session which looked at potential uses of the parser, how it could be developed further and how it could be integrated with other services.

Interfaces and integration

A substantial degree of effort has been aimed at supporting standard protocol access to the gazetteer in order to facilitate machine2machine interfaces. The use of these protocols has been illustrated in both the use of geoXwalk by it's sister project, 'Go-Geo!' and in the Common Information Environment demonstrator 'Heirport'. A simple web based interface for reference access to the gazetteer and which illustrates key functionality offered by geoXwalk has also been implemented. As previously mentioned, a web based interface to the geoparser has been developed.

Users provided feedback on the current geoXwalk demonstrator during the stakeholder workshop on 30 January 2003. Evaluation of geoXwalk was conducted via questionnaires and focus group discussion sessions to allow users to provide feedback on the current design and the future focus of geoXwalk. This included:

- Evaluation of Interface design and potential functionality of both the gazetteer and the geoparser following demonstrations of geoXwalk at the stakeholder workshop 30 January 2003;
- Discussion by workshop participants on how geoXwalk could be integrated with existing services, highlighting any technical and organisational issues which may occur.

Data issues

Significant effort as part of the gazetteer population phase of the project has addressed issues related to improving and enhancing the spatial accuracy and coverage of extant gazetteer sources. Key Ordnance Survey products have been hybridised to produce an enhanced product that is better suited to the demands of spatial querying (as opposed to the purposes of cartographic representation). These are detailed more fully in a separate report: 'Alexandria Digital Library Content Standard and Data Loading & Processing Issues Report.'

Licensing and issues of cost have also been addressed during the project. A variety of aspects remain unresolved in relation to both HE/FE and non-HE/FE access to geoXwalk and Phase III will seek to definitively clarify the terms and conditions of use and their respective costs. Negotiations to this end may require more direct involvement from JISC.

Data licensing for the core datasets that have seeded the demonstrator have benefited from existing licensing arrangements, specifically the Digimap OS agreement and the ESRC's

census programme. While this is currently acceptable within the terms and conditions of usage, longer term use within a shared services environment would require sanction from the key data providers. Discussions to this end have been initiated as part of the project but have not yet reached a clear conclusion. Should geoXwalk be extended beyond HE/FE as part of e.g. a Common Information Environment then the terms and conditions would obviously require revisiting.

As the situation with respect to data licensing is dynamic (witness the role of 'click use' licensing) and the role of shared services within the JISC IE still evolving, it is difficult to definitively negotiate licensing terms. What is clear, is that purchase of additional datasets (if that were required), would be expensive – indeed without the leverage afforded by the Memorandum of Understanding that EDINA have brokered with OS, much of the potential afforded by geoXwalk would be prohibitively expensive for UK academia.

Data issues were discussed at the January workshop through focus group discussion sessions and included data management issues and ideas to handle problems such as loosely defined place names.

Technical investigations

A range of technical areas were investigated as part of the project, including the potential of SOAP and of using Z39.50 as the query protocol for geoXwalk. These are covered in a separate report – 'Technical Investigative Aspects of the geoXwalk Phase II project.'. Essentially, our experiments with SOAP suggested that there were disadvantages from the projects perspective in deploying SOAP based clients and that an XML based mechanism that will permit adoption of a SOAP wrapper upon maturation of the standards, would be a more pragmatic approach. Our experiments with Z39.50 and access to the geoXwalk gazetteer, whilst illustrating that this is possible, remains in our opinion somewhat unnecessarily convoluted and does not permit expression of the full geographic searching potential afforded by geoXwalk. Our conclusion is that the ADL Query Protocol, which is XML based, remains a more tractable protocol for gazetteer querying.

Outputs

By the end of the project the following outputs had been delivered:

- Development of the geoXwalk and geoparser demonstrators
- Development of underpinning software and web interfaces
- Development of methodologies for derivation of gazetteer content and ancillary software to support same
- Creation and ongoing maintenance of comprehensive gazetteer database
- Production of reports including:
    - An Evaluation report;
    - A report on the ADL content standard and data loading and processing issues;
    - A report on the ADL Feature Type Thesaurus and applicability to UK HE/FE
    - A Report detailing Investigative aspects of the project;
- the geoXwalk project website (www.geoxwalk.ac.uk) was created and will continue to be maintained and updated

Impacts

A general rule of thumb is that at least 80% of all information contains some form of spatial reference (source: Association for Geographic Information). Many of the resources within the JISC IE are either **explicitly** or **implicitly** geo-referenced. geoXwalk has the potential, as a shared service, to make geographic based searching as common within the JISC IE as the more traditional methods of searching.

Geography can act as the lowest common denominator to perform a search (most queries will directly or indirectly relate to a specific place or area), the ability that geoXwalk affords to

'crosswalk' across disparate geographies provides an extremely powerful means of searching for and locating resources. Even for those resources that are implicitly geo-referenced, typically just place names embedded somewhere within a record, geoXwalk has identified ways to enhance the spatial searching of such resources. However, co-ordinate based geo-referencing is preferred since this allow places to be represented by the appropriate geographic footprint (settlements as areas; roads, rivers as lines) and offer **persistence,** regardless of name, political boundary or other changes and a **consistent** and powerful framework for spatial searching.

Recognising this, the geo-parser, developed as a part of the geoXwalk project, provides the ability to explicitly geo-reference existing implicitly geo-referenced resources. That is, documents containing place names may be parsed, referenced against the gazetteer and grid references assigned to them. Clearly, the facility to use geography as an additional search filter within the JISC IE is very powerful (note that existing geographic searching is limited because of the paucity and heterogeneity of geographies deployed in the metadata) - geoXwalk overcomes this shortcoming by providing the capacity to be 'geographically agnostic'. The benefits of the geoXwalk approach is clearly demonstrated in it's sister project, Go-Geo!, which is building a geospatial resource discovery tool for the UK tertiary education community. In addition, geoXwalk has been used to extend access to the Statistical Accounts of Scotland as part of the Common Information Environment demonstrator. This clearly illustrates the potential for geoXwalk beyond the JISC IE.

Future priorities for area

A number of outstanding technical issues which have emerged as significant over the course of the project require addressing:

Extending the database to include historical data.

The geoXwalk gazetteer has been populated from 'near contemporary' data sources which enables the demonstrator to resolve queries relating to 'modern' geographies. However, to be of greater utility the database needs to be extended to incorporate 'historical' geographies. *This issue is becoming increasingly critical since any updates of data held by the geoXwalk data means that the existing data becomes historic.*

Adding historical geographies poses a technical challenge in that it becomes imperative to discriminate between existing and new features being added to the database, particularly where data come from different sources and their pedigree varies. Work requires undertaking on developing robust and workable methods to automatically resolve features based on a combination of their type, name and geographical extent/location. This is a non-trivial problem, sometimes referred to as 'map conflation' that requires a pragmatic solution if the gazetteer is to be of practical use. Alongside this, we need to investigate how the historic lineage of a feature is modelled within the database. For example, over time a hamlet may become a village, then a town, and then a city. Is such a place stored as separate entries, one for each type of feature or as a single entry which has associated with it multiple, time-stamped, feature type attributes? The choice of data model adopted will have a resultant impact upon the type of spatial/non-spatial queries capable of being supported.

Performance

The issue of performance needs to be properly investigated (investigations are being conducted under the current Phase 2 project but a more formal and fuller investigation is required for the purposes of supporting a shared service). As a shared service, it is essential that other services using the geoXwalk server do not suffer a performance penalty in terms of query response times because they are waiting on a response from the geoXwalk server. It is important to investigate performance more formally under various load and technical infrastructure models so as to ascertain the optimal cost effective solution that will best service the needs of the JISC IE. If the issue of performance (scalability and response) is not

dealt with before or early on before transition, use of the service by other JISC services may have to be limited to one or two JISC services only.

Trailing to Service

Additionally, the key area of 3rd party integration will be addressed. Phase II identified potential integration targets at the Data Archive, SCRAN and ADS but due to resource scheduling issues at these sites it was not possible to complete the integration work within the Phase II lifetime. The bulk of this workhas had to be deferred to Phase III (see accompanying End of Project Report) and we wish to repurpose the Phase II budget allocation to fund the integration work (£2.5K of this is already committed to SCRAN). Note that Phase III funding lacks any explicit budget for 3rd party integration and we see the requirement to build functioning 'real world' clients with 3rd parties as critical in proving and evaluating the worth of geoXwalk as a potential shared service.